BEST LINEAR PREDICTOR WITH MISSING RESPONSE: LOCALLY ROBUST APPROACH

VICTOR CHERNOZHUKOV MASSACHUSETTS INSTITUTE OF TECHNOLOGY, VIRA SEMENOVA

Abstract. This paper provides asymptotic theory for Inverse Probability Weighing (IPW) and Locally Robust Estimator (LRE) of Best Linear Predictor where the response missing at random (MAR), but not completely at random (MCAR). We relax previous assumptions in the literature about the first-step nonparametric components, requiring only their mean square convergence. This relaxation allows to use a wider class of machine leaning methods for the first-step, such as lasso [12]. For a generic first-step, IPW incurs a first-order bias unless the model it approximates is truly linear in the predictors. In contrast, LRE remains first-order unbiased provided one can estimate the conditional expectation of the response with sufficient accuracy. An additional novelty is allowing the dimension of Best Linear Predictor to grow with sample size. These relaxations are important for estimation of best linear predictor of teacher-specific ([17]) and hospital-specific([18]) effects with large number of individuals.

1. Introduction

Different estimators of missing data models have agreed on the nonparametric way to estimate the probability of missingness. This agreement comes from the intention to impose minimal assumptions on this probability of missingness, called the propensity score, which is typically known little about. Robustness to such assumptions is especially important for the estimation of Best Linear Predictor, a model-free object that is chosen with an intention to assume little known about an economic model. Although large sample results for many missing data estimators have been established, all of them are contingent upon the specific choice of the propensity score plug-in estimators and associated assumptions . We wonder whether it is possible to use conventional missing data estimators under weaker assumptions.

This paper establishes large sample results for Inverse Probability Weighting and Locally Robust Estimator of Best Linear Predictor (BLP) under the assumption of mean square convergence of the first stage estimators. To provide an example where they hold, we establish these results in a high-dimensional sparse linear (logistic) model for the regression function (propensity score). We allow the dimension of BLP to grow with sample size. This result can be applied for estimation best linear predictors of teachers' effectiveness ([17]), hospitals' quality ([18]), and other individual effects with large number of individuals.

Date: February 22, 2017.
Key words and phrases. inference after model selection, locally robust estimation, inverse probability weighting,propensity score.

This paper builds on earlier work. Large sample results for Inverse Probability Weighting (IPW) and Locally Robust Estimator (LRE) have been established in [5], [3], [11], [9], [16]. IPW in [3] approximates the propensity score by logistic polynomial function and exploits the local robustness property of the polynomial estimators to establish first-order unbiasedness. For generic propensity score estimators, IPW may incur first-order bias unless the regression function is almost linear in the predictors. We establish a condition on the size of approximation error which results in first-order zero bias of IPW.

To avoid generic first-order bias of IPW, one can use efficient influence function of [4] that defines LRE. In order to apply it, one have to estimate the propensity score and the regression function at the first step. A version of LRE ([11]), called Semiparametric Doubly Robust Estimator, uses local probit regression for the propensity score and local linear regression for the regression function. Another version of LRE ([16]), called Inverse Probability Tilting, suggests estimating the propensity score parameter and BLP jointly, by solving a system of efficient moment equations. However, all these large sample results are contingent upon the nonparametric estimators they use. Moreover, they require additional tuning, such as the choice of kernel bandwidth in [11] or specifying a dictionary of approximating functions for the propensity score and the regression function [16]. Therefore, with the goal of being oblivious to the first-stage estimators, we will start with efficient influence function of [4] and pursue locally robust approach of [9].

This paper has several contributions. First, the mean square convergence condition of the nonparametric part is weaker than previously imposed. Second, we derive large sample results of IPW and LRE in a high-dimensional sparse model which cannot be analyzed with previous nonparametric estimators. Finally, we allow the dimension of Best Linear Predictor to grow with the sample size and the lower bound of the propensity score to vanish to zero, allowing the total sample size to grow as a degree-2-polynomial of the complete sample size.

The content of the paper is as follows. Section 2 defines the semiparametric missing data model for best linear predictor and its estimators. Section 3 establishes asymptotic normality results of Inverse Probability Weighting and Locally Robust Estimator under mean square convergence assumptions of the propensity score and regression function. Section 4.1 justifies the assumptions of Section 3 a high-dimensional sparse linear (logistic) model for the regression function (propensity score). We provide Monte Carlo comparison of Inverse Probability Weighting and Locally Robust Estimator in Section 4.3.

2. Set-Up

Here we introduce a model for best linear predictor with missing response and the estimators of it. The model is similar to those in [11] and [16]. The different from ([11], [16]) is that we allow the dimension of the best linear predictor to grow with the sample size. Also, the lower uniform bound on the propensity score $\underline{s}s_{n,N}$ is allowed to vanish to zero. We believe we are the first to offer these technical relaxations in missing data literature.

2.1. **Model for BLP with Missing Response.** Let $\{(\Delta_i, x_i, \Delta_i y_i)_{i=1}^N\}$ be the observed data, where y_i is a response variable, $x_i \in R^p$ is a vector of regressors, $\Delta_i \in \{1, 0\}$ is an indicator of y_i being observed. We refer to an observation (x_i, y_i) as a **complete** one, and to an observation $(x_i, 0)$ as an **incomplete** one. Let $x \to g_N(x)$ be the conditional mean of the outcome y_i, which we refer to as regression function, and $s \to s_N(x)$ be the propensity score. The regression function and the propensity score are allowed to change N, but we will suppress the dependence in further notation.

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, ..., N \tag{2.1}$$

$$s(x_i) : = \mathbb{E}\big[\Delta_i = 1 | x_i\big], \quad g(x_i) := \mathbb{E}\big[y_i | x_i\big] \tag{2.2}$$

$$s_i : = s(x_i), \quad g_i := g(x_i) \tag{2.3}$$

Let $p_i \in R^k$ be a given k-dimensional dictionary of regressors with respect to x_i, that is

$$p_i := p(x_i)$$

In making asymptotic statements, we assume that the number of all observations $N \to \infty$, the number of complete observations $n = \sum_{i=1}^N \Delta_i \to \infty$, the dimension of original and technical regressors $p = p_n \to \infty$ and $k = k_n \to \infty$, respectively.

The object of interest is **Best Linear Predictor** with respect to the regressors $p(x_i) = (p_1(x_i), ..., p_k(x_i))'$:

Definition 2.1 (Best linear predictor (BLP))**.**

$$\beta := \arg\min_{b \in R^k} \mathbb{E}(g(x_i) - p_i^\top b)^2 = \arg\min_{b \in R^k} \mathbb{E}(y_i - p_i^\top b)^2, \tag{2.4}$$

Given the regression function $g(x_i)$ and BLP β, define the approximation error of β as follows:

$$\forall x \in \mathcal{X}, r(x) := r_g(x) := g(x) - p(x)^\top \beta$$

$$r_i := r(x_i) = g(x_i) - p(x_i)^\top \beta$$

Assumption 2.1 (Missingness at random)**.**

$$\mathrm{P}(\Delta_i = 1 | x_i = x, y_i) = \mathrm{P}(\Delta_i = 1 | x_i = x) = s(x), x \in \mathcal{X}$$

Let $\zeta_i := \frac{\Delta_i}{s(x_i)} - 1$. By Assumption 2.1, it is uncorrelated with any function of x.

Assumption 2.2 (Strong Overlap of Normalized Propensity Score)**.**

$$\exists s_{n,N}, \underline{s}, \bar{s} : 0 < \underline{s} < \frac{s(x)}{s_{n,N}} < \bar{s} < 1$$

We refer to the model defined by 2.2- 2.3 as a BLP with Missing Response. Similar models have been studied in [4] and [16], with an exception that these papers have assumed finite-dimensional parameter of interest and had their propensity score uniformly bounded from zero.

2.2. **Estimators.** Here we will describe the general form of IPW and LRE estimators, without specifying the procedure for their nonparametric components. Both of them have two stages, performed on different samples. In the first stage, we estimate the propensity score and the regression function. To get IPW, we reweigh the complete data by inverse propensity score and compute the linear projection of the reweighed response on the reweighed covariates. To get LRE, we (1) impute zeros instead of missing response, (2) reweigh the resulting responses by inverse propensity score, (3) subtract the regression function reweighed by observation regression error ζ_i and (4) compute the linear projection of the difference on the covariates.

In what follows below, we split the sample into the two parts at random, and call them a training one(a) and estimation one (b). On the training sample, we estimate the nuisance functions $g(x), s(x)$. On the estimation sample, we compute the estimator $\widehat{\beta}_{b,a}$. Then, we flip the roles of the parts a, b and construct $\widehat{\beta}_{a,b}$. The final estimator $\widehat{\beta} = \frac{\widehat{\beta}_{b,a}+\widehat{\beta}_{a,b}}{2}$. We call this procedure cross-fitting.

Definition 2.2 (Inverse Probability Weighting). *Let $\widehat{s}_i := \widehat{s}(x_i) = \widehat{s}_a(x_i)$ be an estimator of the propensity score $s(x)$. Define the inverse probability weighting as follows:*

$$\widehat{\beta}_{IPW} := (\sum_{i\in b} p_i p_i' \frac{\Delta_i}{\widehat{s}_i})^{-1} \sum_{i \in b} p_i y_i \frac{\Delta_i}{\widehat{s}_i}$$

Definition 2.3 (Locally Robust Estimator: General Case). *Let $\widehat{s}_i := \widehat{s}(x_i) = \widehat{s}_a(x_i), \widehat{g}_i = \widehat{g}(x_i) = \widehat{g}_a(x_i)$ be a nonparametric estimator of the propensity score $s(x)$ and regression function $g(x)$, respectively. Define locally robust least squares as follows:*

$$\tilde{y}_i := \widehat{g}_i + \frac{\Delta_i}{\widehat{s}_i}(y_i - \widehat{g}_i), i \in b \tag{2.5}$$

$$\widehat{\beta}_{LR} := (\sum_{i\in b} p_i p_i')^{-1} \sum_{i\in b} p_i \tilde{y}_i \tag{2.6}$$

Sample splitting automatically sets the mean of the interaction of the first-stage estimation error with the second-stage sampling error to zero. This allows to use a very broad set of ML predictive methods to estimate the nuisance parameter, such as lasso, random forests, deep neural nets, boosted trees, as well as various hybrids and aggregates of these methods. There is no efficiency loss due to cross-fitting (Theorem 3.6).

3. Asymptotic Theory for IPW and LRE

In this section we provide asymptotic normality results for IPW and LRE under the assumptions of the propensity score and regression functions estimated at rate $o_P(n^{-1/4})$. Unlike previous versions IPW and LRE, this assumption does not impose any structure on these functions. We show that IPW is valid for inference if the Best Linear Predictor is close to the regression function. Otherwise, it incurs a first-order bias. In contrast, LRE avoids this bias by exploiting orthogonality of the first-step estimation error to second-stage observation regression error and response regression error. Therefore, it remains valid regardless of BLP approximation quality.

Assumption 3.1 (Stability)**.** *The matrix* $\mathbb{E}p_i p_i'$ *has its minimal eigenvalue* $\underline{\lambda}_n > \underline{\lambda} > 0$ *bounded away from zero.*

Let $\xi_k = \sup_{x \in \mathcal{X}} \|p(x)\|_2^2$ be a bound on the norm of the technical regressors. For example, $\xi_k \lesssim k$ for Legendre polynomial series, $\xi_k \lesssim \sqrt{k}$ for Fourier series, B-spline series, and local polynomial partition series. To ensure that empirical design matrix converges to its expectation, we assume that the regressors grow sufficiently slow:

Assumption 3.2 (Bounded covariates)**.**

$$\frac{\xi_k^2 \log N}{N s_{n,N}} \to 0$$

or, equivalently,

$$\frac{\xi_k^2 \log N}{n} \to 0$$

Let the sequences of the finite constants c_k and $l_k c_k$ denote the upper bound on the second and infinity norm of the misspecification error, respectively. Namely, let c_k and $l_k c_k$ be such that $f \in \mathcal{G}$, $\|r\|_{F,2} := \sqrt{\int r^2(x) dF(x)} \leqslant c_k$ and $\|r\|_{F,\infty} := \sup_{x \in \mathcal{X}} |r(x)| \leqslant l_k c_k$. We define the regime of small misspecification as

Assumption 3.3 (Small Misspecification)**.** $\sqrt{n}\boldsymbol{ps}_N (l_k c_k) \frac{1}{\sqrt{s_{n,N}}} \to 0$*, or, equivalently,* $\sqrt{N}\boldsymbol{ps}_N (l_k c_k) \to 0$*,*

Assumption 3.3 provides middle ground requirement between two knife-edge cases. In case of exact linear specification $l_k c_k = 0$, the IPW becomes weighted least squares in a true linear model, and is consistent $\sqrt{n}$ asymptotically normal regardless of the weights. In case knowledge of the true propensity score $\mathbf{ps}_N = 0$, IPW is unbiased. This middle ground requirement means that including sufficiently many regressors $p(x)$ for the Best Linear Predictor and having sufficiently good propensity score results in valid inference, and is more realistic than any of the knife-edge cases per se.

Let $\mathbf{ps}_N$, $\mathbf{g}_n$ be the rate of convergence of the estimators of the normalized propensity score and the regression function, respectively. Namely, let

Assumption 3.4 (High-Level Assumptions on First Stage Estimation)**.**

$$\frac{1}{s_{n,N}}\sqrt{\mathbb{E}(s(x)-\widehat{s}_a(x))^2} = o_P(\boldsymbol{ps}_N) \lesssim n^{-1/4} \tag{3.1}$$

$$\sqrt{\mathbb{E}(g(x)-\widehat{g}_a(x))^2} = o_P(\boldsymbol{g}_n) =\lesssim n^{-1/4} \tag{3.2}$$

Assumption 3.4 requires $o_P(n^{-1/4})$ mean square convergence of the normalized propensity score and the regression function. Since the response variable is partially missing, any standard convergence result with no missingness case has to be justified anew. Missing at Random (2.1) assumption partially helps with that: $\mathbb{E}(\mathbb{E}\big[y_i|x_i\big]|\Delta_i = 1) = \mathbb{E}\big[y_i|x_i\big]$, which implies that conditional density $f(y|x)$ can be estimated consistently on the complete sample only $(\Delta_i x_i, \Delta_i y_i)_{i=1}^n$. This is sufficient for consistency of local linear estimator used by [11] . However, other nonparametric estimators may impose an assumption of the distribution of observed covariates $((\Delta_i x_i)_{i=1}^N$, and the argument above is insufficient. We give an example of such setting in Section 4.1.

Theorem 3.1. *(L^2 rate of the IPW) Let the Assumptions 3.1, 3.2 hold. Then,*

$$\|p^\top(\widehat{\beta}_{IPW}-\beta)\|_{F,2} \lesssim_P \sqrt{k/n}(1+l_k c_k)(1+\sqrt{k/s_{n,N}}\boldsymbol{ps}_N)$$

This is our first main result in this paper. $\widehat{\beta}_{IPW}$ has the same convergence rate as regular ordinary least squares estimator, namely $\sqrt{k/n}(1+l_k c_k)$, provided the $\sqrt{k/s_{n,N}}\mathbf{ps}_N$ is small.

Theorem 3.2 (Asymptotic Linearity of $\widehat{\beta}_{IPW}$)**.** *Let the Assumptions 3.1, 3.2,3.3, be satisfied. Then, for any $\alpha \in S^{k-1}$,*

$$\sqrt{n}\alpha^\top(\widehat{\beta}_{IPW}-\beta) = \alpha^\top Q^{-1} G_n[p_i(\epsilon_i + r_i)\frac{\Delta_i}{s_i}] + R_{1,n}(\alpha), \tag{3.3}$$

where the term $R_{1,n}(\alpha)$ obeys:

$$R_{1,n}(\alpha) \lesssim_P n^{1/4} l_k c_k + n^{-1/2}$$

Let $\Omega_{IPW} = Q^{-1}\mathbb{E}p_i p_i^\top(\epsilon_i + r_i)^2\frac{\Delta_i}{s_i^2}Q^{-1}\mathbb{E}s_i$

Theorem 3.3 (Asymptotic Normality of $\widehat{\beta}_{IPW}$)**.** *Let the Assumptions 3.1, 3.2,3.3, be satisfied, and $R_{1,n}(\alpha)$, defined in 3.3, obeys $R_{1,n}(\alpha) = o_P(1)$. In addition, assume that*

$$\sup_{x\in\mathcal{X}} \mathbb{E}\epsilon_i^2 1\{|\epsilon_i| > M\}|x_i = x \to 0, M \to \infty \tag{3.4}$$

uniformly over n. Then, for any $\alpha \in S^{k-1}$,

$$\sqrt{n}\frac{\alpha^\top(\widehat{\beta}_{IPW} - \beta)}{\|\alpha^\top \Omega^{1/2}\|} =_d N(0,1)$$

Moreover, $\widehat{\Omega}_{IPW} = (\mathbb{E}_N p_i p_i^\top)^{-1}\mathbb{E}_N p_i p_i^\top (y_i - p_i'\widehat{\beta})^2 \frac{\Delta_i}{\widehat{s}_i^2}(\mathbb{E}_N p_i p_i^\top)^{-1}\frac{n}{N}$ is a consistent estimator of Ω_{IPW}.

Theorem 3.3 is our second main result in the paper. Under Assumption 3.3 about small misspecification of BLP, it establishes $\sqrt{n}$ asymptotic normality of IPW and consistency of a simple sample analog of the variance estimator.

In case Assumption 3.3 does not hold, generic propensity score estimators may introduce first-order bias in IPW. In an exceptional case, IPW in [3] this bias is numerically zero by virtue of local robustness of polynomial estimators. However, this property may not hold for generic estimator of the propensity score.

Sensitivity of IPW to the misspecification of linear model can be explained as follows. For example, let the covariates' space consist of the single variable age. The response variable is log wages which we believe is a quadratic function of age with an added sampling error. A propensity score estimator such as kernel or local linear will have larger estimation error on the tail of the age distribution compared to its center, if there are more observations there. At the same time, the approximation error would be larger on the upper tail of the age distribution, by virtue of a quadratic function. Simultaneous presence of approximation error and the estimation error of the propensity score creates the first-order bias.

Theorem 3.4 (Asymptotic Linearity of $\widehat{\beta}_{LR}$)**.** *Under the Assumptions 3.1, 3.2, for any $\alpha \in S^{k-1}$,*

$$\sqrt{n}\alpha^\top(\widehat{\beta}_{LR} - \beta) = \sqrt{n}/\sqrt{n}\alpha^\top Q^{-1}\mathbb{G}_n[p_i(\frac{\Delta_i \epsilon_i}{s_i} + r_i)] + R_{1n}(\alpha),$$

where $R_{1n}(\alpha) = o_P(n^{-1/2})$

Define $\Omega_{LR} := \mathbb{E}s_i Q^{-1}\mathbb{E}[p_i p_i^\top(\frac{\Delta_i \epsilon_i}{s_i} + r_i)^2]Q^{-1}$

Theorem 3.5 (Asymptotic Normality of $\widehat{\beta}_{LR}$)**.** *Assume $\sup_{x \in \mathcal{X}} \mathbb{E}\epsilon_i^2 1\{|\epsilon_i| > M\}|x_i = x \to 0, M \to \infty$ uniformly over n. Under the Assumptions 3.1, 3.2, and $R_{1n}(\alpha) = o_P(n^{-1/2})$ for any $\alpha \in S^{k-1}$,*

$$\sqrt{n}\frac{\alpha^\top(\widehat{\beta}_{LR} - \beta)}{\|\alpha^\top \Omega_{LR}^{1/2}\|} =_d N(0,1)$$

Let the estimate of the noise be $\widehat{\epsilon}_i := y_i - \widehat{g}_i, \forall i : \Delta_i = 1$ and of the approximation error be $\widehat{r}_i := \widehat{g}_i - p_i^\top \widehat{\beta}_{LR}$. Then, a consistent estimator of $\widehat{\Omega}_{LR}$ can be obtained as $\widehat{\Omega}_{LR} = \frac{n}{N}(\mathbb{E}_N p_i p_i^\top)^{-1}\mathbb{E}_N p_i p_i^\top \left[(\frac{\Delta_i \widehat{\epsilon}_i}{\widehat{s}_i} + \widehat{r}_i)^2\right](\mathbb{E}_N p_i p_i^\top)^{-1}$

Theorem 3.6 (No efficiency loss due to sample splitting)**.**

Theorem 3.5 is our third main result in the paper. It establishes $\sqrt{n}$ asymptotic normality of LRE and consistency of a simple sample analog of the variance estimator.

Theorems 3.3 and 3.5 provide an improvement over the prior assumptions on the first-stage estimators. For example, the asymptotic normality of the Doubly Robust Estimator SDRE of [11] utilizes local polynomial estimator and relied on its U-statistic property. Asymptotic result for IPW [3] used logistic polynomial estimators and assumed continuous differentiability of the propensity score . In addition to these restrictive assumptions, these nonparametric estimators require additional tuning parameters to be set. In contrast, the rate requirement used here is weaker and is satisfied by machine learning methods, such as lasso [12] and random forests.

Although efficiency is beyond the scope of this paper, we can make several observations regarding precision properties of LRE. First, its variance coincides with the information matrix, provided by [4], which is equal to $\mathbb{E}s_i Q^{-1}\mathbb{E}[p_i p_i^\top \left(\frac{\Delta_i \epsilon_i}{s_i} + r_i\right)^2]Q^{-1}$. Second, Theorem 3.6 states that cross-fitting fully retains the efficiency lost incurred due to sample splitting. Finally, LRE is locally efficient at the restriction $\mathrm{Var}(y_i|x_i) = \sigma_\epsilon^2 = const.$

4. Asymptotic Theory for High Dimensional Sparse Model

To give an example of environment where the mean square convergence assumption (3.4) of Section 3 hold, we consider a high-dimensional sparse linear (logistic) model for the regression function (propensity score). We have chosen it for three reasons. First, the earlier specifications of IPW ([3]) and LRE ([11], [16]) cannot be applied to it due to dimensionality restrictions. Second, unlike the low-dimensional setting, the need in the reweighing of the covariates appears in the high-dimensional one for consistency of regression function estimator. Finally, the high-dimensional sparse setup is appropriate for analyzing modern high-dimensional datasets.

The notation for this section is as follows. Let the regression function be:

$$g(x_i) = x_i'\eta + c_{s_\eta}$$

and the propensity score be:

$$s(x_i) = L(x_i'\gamma),$$

where $\eta \in R^p$ and $\gamma \in R^p$ are sparse vectors with the sparsity indices $s_\eta = \|\eta\|_0$ and $s_\gamma = \|\gamma\|_0$, c_{s_η} is the approximation error, and $L(t) = \frac{\exp t}{1+\exp t}$ is the logistic function.

4.1. **Logistic Model for the Propensity Score.** Here we derive the L^2 convergence of the propensity score. The proof is essentially based on [13].

Assumption 4.1 (Assumptions on the Covariates for Penalized Logistic Regression)**.** *Let* $w_i = L(x_i'\gamma)(1-L(x_i'\gamma))$ *be the derivative of the logistic function at* $x_i'\gamma$*. The weights* w_i *satisfy* $\min_{i\leqslant N} w_i \geqslant c > 0$ *with probability* $1-\Delta_N$*. Denote the weighted restricted eigenvalue as follows: For a given* $\bar{c} \geqslant 0$*,*

$$\kappa(\bar{c}) := \min_{\|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \delta\neq 0} \frac{\sqrt{s}\|\sqrt{w_i}\tilde{x_i}'\delta\|_{2,n}}{\|\delta_T\|_1} > 0$$

For a subset $A \subset R^p$ *let the non-linear impact coefficient be:*

$$\bar{q}_A = \inf_{\delta\in A} \mathbb{E}_n[w_i|\tilde{x_i}'\delta|]^{3/2}/\mathbb{E}_n w_i|\tilde{x_i}'\delta|^3$$

Denote the minimal and maximal restricted eigenvalues be:

$$\psi_{(r)}(\boldsymbol{c}) := \min_{\|\delta_{T^c}\|_1 \leqslant \boldsymbol{c}\|\delta\|_1} \frac{\|\sqrt{w_i}\tilde{x_i}'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}}$$

and

$$\psi_{(r)}(\boldsymbol{c}) := \min_{1\leqslant\|\delta\|_0\leqslant m} \frac{\|\sqrt{w_i}\tilde{x_i}'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}}$$

Assume that there exists $s = s_\gamma$ *such that* $\|\gamma_0\|_0 \leqslant s_\gamma$ *. The sparse minimal and maximal eigenvalues are bounded* $c \leqslant \phi_{min}(sl_n) \leqslant \phi_{max}(sl_n) \leqslant C$ *with probability* $1-\delta_n$*. The sparsity index* s_γ *and overall number of covariates* p *obey the following growth conditions* $K_x^2 s_\gamma^2 \log^3(p,n) \leqslant n\delta_n$ *and* $K_x^q \log\min(p,n) \leqslant n\delta_n$ *for some fixed* $q > 4$*. Let* $\bar{q}_{\Delta_c} > 3(1+\frac{1}{c})\lambda\sqrt{s_\gamma}/n\kappa_c$*.*

The Assumption 4.1, based on the Condition L and Lemma 13 of [13], is needed for the consistency of the logistic lasso estimator. Their proof exploits the convexity of the logistic penalty 4.1 to show that close values of the objective function imply close values of the parameters in $\|\|_{2,n}$ norm with the weighted regressors $\sqrt{w_i}x_i$. Next, assuming that $\min_{i\leqslant N} w_i \geqslant c > 0$, Lemma 7 is used to establish the equivalence between restricted sparse eigenvalues of weighted $\sqrt{w_i}x_i$ and unweighted x_i covariates. Finally, assuming standard restricted eigenvalue assumptions $c \leqslant \phi_{min}(sl_n) \leqslant \phi_{max}(sl_n) \leqslant C$ with probability $1-\delta_n$, the convergence in l_1 norm of the parameters is established.

Definition 4.1 (Logistic Lasso Estimator for the Propensity Score)**.** *For the level of penalty* $\lambda_\gamma = \frac{1.1}{\sqrt{n}}\Phi^{-1}(1-0.05/\max(n, p\log n))$*, let the estimator of the propensity score be defined as logistic lasso:*

$$\Lambda(\gamma) = \sum_{i\in a} \log(1+\exp(x_i'\gamma)) - \Delta_i x_i \tag{4.1}$$

$$\widehat{\gamma} := \arg\min_{\gamma\in R^p} \Lambda(\gamma) + \lambda_\gamma\|\gamma\|_1 \tag{4.2}$$

$$\widehat{s}_i := \widehat{s}(x_i) = L(x_i'\widehat{\gamma}), i \in b \tag{4.3}$$

Lemma 4.1 (Consistency of Logistic Lasso)**.** *Let the nuisance parameter γ satisfy Assumption 4.1. Let the penalty level $\lambda_\gamma = \frac{1.1}{\sqrt{n}}\Phi^{-1}(1-0.05/\max(n, p\log n))$ Then,*

$$\|\widehat{s}(x) - s(x)\|_{L^2} \lesssim_P \frac{\lambda_\gamma s_\gamma}{n\kappa_c^2} \lesssim_P n^{-1/4}$$

The proof of Lemma 4.1 is is a direct corollary of the convergence of $\|\widehat{\gamma} - \gamma\|_1$ Lemma 13 in [13]. By the mean value theorem,

$$\forall x \quad \|\widehat{s}(x) - s(x)\| \leqslant \max_{x\in\mathcal{X}} L(x'\gamma)(1 - L(x'\gamma)) \max_{x\in\mathcal{X}} \|x\|_\infty \|\widehat{\gamma} - \gamma\|_1$$

which implies the mean square convergence.

4.2. **Linear Model for Regression Function.** Since the response variable is missing, the convergence result of [12] for the regression function has to be justified anew. Since the sampling error is conditionally independent from x, any transformation of the covariates leaves the gradient mean zero. Reweighing by the inverse propensity score leaves the mean of the covariates unchanged and therefore does not require additional assumptions on the propensity score.

Assumption 4.2 (Assumptions on the Covariates for Penalized Linear Regression)**.** *Let $\tilde{x}_i, i = 1, 2, ..., N$ are i.i.d bounded zero-mean vectors, with $\max_{1\leqslant i\leqslant N}, \max_{1\leqslant j\leqslant p} |\tilde{x}_{ij}| \leqslant K_B$ for all N and p. Assume that the population design matrix $\mathbb{E}\tilde{x}_i\tilde{x}_i'$ has ones on the diagonal, and its $s\log N$ sparse eigenvalues are bounded from above by $\phi < \infty$ and bounded from below by $\kappa^2 > 0$.*

The Assumption 4.2 provides an example of $(x_i)_{i=1}^N$ whose $s\log n$ restricted sparse eigenvalues are bounded from above and below. For the proof of this fact, please see Proposition 2 (ii) of [15].

Definition 4.2 (Lasso Estimator for the Regression Function)**.** *For the level of penalty $\lambda_\eta = \frac{1.1}{\sqrt{n}}\Phi^{-1}(1-0.05/\max(n, p\log n))$, let the estimator of the regression function g be defined as weighted linear lasso:*

$$\widehat{\eta} := \arg\min_{\eta\in R^p} \sum_{i\in a} \Big(\frac{\Delta_i}{\widehat{s}_a(x_i)} y_i - \frac{\Delta_i}{\widehat{s}_a(x_i)} x_i'\eta\Big)^2 + \lambda_\eta\|\eta\|_1 \tag{4.4}$$

$$\widehat{g}_i = x_i'\widehat{\eta}, i \in b \tag{4.5}$$

Lemma 4.2 (Consistency of Lasso on Complete Data)**.** *Let Assumptions 4.1 and 4.2 hold. Let $\lambda_\eta \lesssim_P \sqrt{\log p/n}$. Then,*

$$\|x_i'(\widehat{\eta} - \eta)\|_{L^2} \lesssim \frac{\lambda_\eta s_\eta}{n\kappa(c)^2} \lesssim_P n^{-1/4}$$

Lemma 4.2 is established in three steps. First, we show that the restricted $s \log n$ sparse eigenvalues of the matrix $\mathbb{E}_N \frac{\Delta_i}{\widehat{s}_a^2(x_i)} x_i x_i'$ with estimated weights are $o_P(n^{-1/4})$ close to those of the matrix $\mathbb{E}_N \frac{\Delta_i}{s^2(x_i)} x_i x_i'$ with the true weights. Next, $(\frac{\Delta_i}{s(x_i)} x_i)_{i=1}^N$ is an i.i.d. mean zero sequence with its restricted sparse $s \log n$ eigenvalues bounded from above and below ([15]). Therefore, reweighing by the estimate of the propensity score has not introduced additional assumptions about it. Second, the mean of the noise $\mathbb{E}_N \frac{\Delta_i}{\widehat{s}_a(x_i)} x_i \epsilon_i$ is not affected by reweighing since the sampling error is independent of covariates. Hence, lasso consistency in l_1 norm is proved the same way as in Theorem 1 in [12]. Finally,

$$\forall x \quad |g(x) - \widehat{g}(x)| \leqslant \max_{x \in' \mathcal{X}} \|x\|_\infty \|\widehat{\eta}_a - \eta\|_1 \lesssim_P o_P(n^{-1/4})$$

The proof of Lemma 4.2 suggests an interesting observation about the purpose of the reweighing. As before, any covariate-reweighing of the noise $\mathbb{E}_N x_i \epsilon_i$ does not change its mean. However, covariate-reweighing (including reweighing by identity) affects the *mean* of the covariates. Therefore, additional $s \log n$ restricted eigenvalue assumptions have to be imposed on the propensity score. Inverse-propensity-score weighting is the only reweighing applied to complete data which preserves covariates' mean and therefore does not require any assumptions about the propensity score beyond convergence.

Corollary 4.1 (Asymptotic Normality of IPW:HDS). *Let the nuisance parameter γ satisfy Assumption 4.1 and the sparsity index s_γ satisfies $\frac{s_\gamma \log p \vee n}{n^{-1/4}} = o_p(1)$. Then, by Theorem 4.1, the rate of mean square convergence $\|s(x) - \widehat{s}_a(x)\|_{L^2} = o_p(n^{-1/4})$. Assume small misspecification ($\sqrt{n} l_k c_k n^{-1/4}/s_{n,N}^{3/2}$) and the Lindenberg condition:*

$$\sup_{x \in \mathcal{X}} \mathbb{E}\epsilon_i^2 1\{|\epsilon_i| > M\} | x_i = x \to 0, M \to \infty \tag{4.6}$$

uniformly over n. Then, for any $\alpha \in S^{k-1}$, we have

$$\sqrt{n} \frac{\alpha^\top (\widehat{\beta}_{IPW} - \beta)}{\|\alpha^\top \Omega^{1/2}\|} =_d N(0,1)$$

where $\Omega_{IPW} = Q^{-1} \mathbb{E} p_i p_i^\top (\epsilon_i + r_i)^2 \frac{\Delta_i}{s_i^2} Q^{-1} \mathbb{E} s_i$

Corollary 4.2 (Asymptotic Normality of LR:HDS). *Let the nuisance parameter γ satisfy Assumption 4.1 and the sparsity index s_γ satisfies $\frac{s_\gamma \log p \vee n}{n^{-1/4}} = o_p(1)$. Then, by Theorem 4.1, the rate of convergence of $\|s(x) - \widehat{s}_a(x)\|_{L^2} = o_p(n^{-1/4})$. Let the nuisance parameter η satisfy $\frac{s_\eta \log p}{n} = o_p(n^{-1/4})$. Then, by Theorem 4.2, the rate of convergence of $\|g(x) - \widehat{g}_a(x)\|_{L^2} = o_p(n^{-1/4})$. Under the Lindenberg condition:*

$$\sup_{x\in\mathcal{X}} \mathbb{E}\epsilon_i^2 1\{|\epsilon_i| > M\}|x_i = x \to 0, M \to \infty \tag{4.7}$$

uniformly over n. Then, for any $\alpha \in S^{k-1}$, we have

$$\sqrt{n}\frac{\alpha^\top(\widehat{\beta}_{LR} - \beta)}{\|\alpha^\top \Omega^{1/2}\|} =_d N(0,1)$$

where $\Omega_{LR} := \mathbb{E}s_i Q^{-1}\mathbb{E}[p_i p_i^\top(\frac{\Delta_i \epsilon_i}{s_i} + r_i)^2]Q^{-1}$

Corollaries 4.1 and 4.2 state asymptotic normality of best linear predictor in a high-dimensional sparse model.

4.3. **Monte Carlo Simulations.** Our Monte-Carlo results support theoretical findings. Under the small misspecification, IPW and LRE have similar performance. Otherwise, IPW produces bias and poor coverage while LRE remains valid.

To study the finite sample properties of IPW and LRE, we consider a high-dimensional sparse model. Let the propensity score $s(x)$ be :

$$s(x_i) = \frac{\exp(x_i'\eta)}{\exp(x_i'\eta) + 1}$$

and the regression function be:

$$g(x_i) = x_i'\gamma,$$

where the coefficients η, β, γ exhibit the following declining pattern:

$$\eta = (1, 1/2, 1/3, ..., 1/p_\eta, 0, 0, ..0)', \tag{4.8}$$

$$\gamma = [(1, ..., 1/k), c_\gamma(1/(k+1).., 1/p)]', \tag{4.9}$$

Let the Best Linear Predictor be the projection of y_i on the constant and first k regressors of x_i:

$$y_i = \beta_0 + x_{i,1:k}'\beta + u_i, \quad \mathbb{E}u_i x_{i,1:k} = 0 \tag{4.10}$$

The data generating process is as follows. The covariates $(x_i)_{i=1}^N$ are i.i.d normally distributed: $x_i \sim N(0,\Sigma)$. The covariance matrix is a Toeplitz matrix with correlation coefficient ρ: $\Sigma_{ij} = \rho^{|i-j|}$, ρ will be chosen to govern misspecification size. The variance of the regression noise $\epsilon_i \sim N(0,1)$ is normalized to one.The dimension of the covariates x_i is $p = 300$ in the regression function and $p_\delta = 100$ in the propensity score. The total sample size is $N = 500$, and the size of complete size is $n = 250$. The number of the predictors $k = 6$. The parameters c_γ and ρ determine the specification quality and will be set later.

Table 1 shows the bias, standard deviation, root mean squared error, and rejection frequency for Ordinary Least Squares, Inverse Probability Weighting, and Locally Robust Estimator under

small misspecification. It is introduced by setting the correlation between regressors to be moderate ($\rho = 0.5$) and the omitted regressors to be scaled by a small constant ($c_\gamma = 0.05$). The results in an approximately correct specification of the linear model $R^2 = 0.74$. In that case, all the three estimators have small bias and good coverage property. IPW and LRE have similar performance, since their covariance matrices coincide under correct specification ($\Omega_{IPW} = \Omega_{LRE}$). Since the linear model is close to the true one, OLS is best linear conditionally unbiased estimator, and therefore has smaller variance than IPW.

Table 2 shows finite sample properties of IPW, LRE, and OLS under large misspecification. It is introduced by scaling the omitted regressors by a large constant ($c_\gamma = 6.33$). Since the sampling error variance if fixed, the decrease of R^2 from $R^2 = 0.74$ to $R^2 = 0.1$ indicates larger misspecification. .OLS suffers from selection bias. IPW incurs first-order bias due to the propensity score estimation error. For example, its rejection frequency for the constant coefficient β_0 is 0.201 instead of the nominal 0.05. LRE remains valid.

To conclude, IPW and LRE may be computed simultaneously for robustness check. If their values are far apart, this may be a signal for misspecification, in which case LRE should be chosen for inference.

	OLS	IPW	LR	OLS	IPW	LR	OLS	IPW	LR	OLS	IPW	LR
	Bias			St.Error			RMSE			Rej.Freq.		
1	0.001	0.001	0.006	0.023	0.036	0.035	0.001	0.002	0.006	0.001	0.002	0.002
2	-0.000	0.000	0.004	0.022	0.041	0.041	0.022	0.041	0.041	0.031	0.041	0.036
3	-0.000	-0.001	0.002	0.026	0.045	0.046	0.026	0.045	0.046	0.028	0.032	0.031
4	0.000	-0.001	0.003	0.026	0.044	0.044	0.026	0.044	0.044	0.026	0.032	0.036
5	0.001	0.000	0.003	0.026	0.044	0.044	0.026	0.044	0.044	0.026	0.032	0.032
6	-0.001	-0.001	0.002	0.026	0.044	0.044	0.026	0.044	0.044	0.021	0.033	0.027

Table 1. Bias, St.Error, RMSE,Rejection Frequency of OLS, IPW, LR. Small misspecification. Test size $\alpha = 0.05$.

	OLS	IPW	LR	OLS	IPW	LR	OLS	IPW	LR	OLS	IPW	LR
	Bias			St.Error			RMSE			Rej.Freq.		
1	0.121	0.109	-0.001	0.074	0.116	0.075	0.142	0.159	0.075	0.339	0.201	0.044
2	-0.047	-0.043	0.001	0.075	0.134	0.077	0.088	0.141	0.077	0.096	0.097	0.047
3	-0.015	-0.014	-0.002	0.070	0.110	0.076	0.072	0.110	0.076	0.058	0.057	0.052
4	-0.004	-0.006	-0.002	0.068	0.112	0.074	0.069	0.112	0.074	0.054	0.069	0.048

Table 2. Bias, St.Error, RMSE, Rejection Frequency of OLS, IPW, LR. Large misspecification. Test size $\alpha = 0.05$.

5. APPENDIX

5.1. Notation.

5.1.1. *Overall Notation.* Throughout the paper, the symbols P and $\mathbb{E}$ denote probability and expectation operators with respect to a generic probability measure. $\mathbb{E}_N, \mathbb{E}_n, \mathbb{E}_{n,a}, \mathbb{E}_{n,b}$ denote the sample average, taken in the whole sample ($\mathbb{E}_N$), the sample of complete observations ($\mathbb{E}_n$), training sample ($\mathbb{E}_{n,a}$), and main sample ($\mathbb{E}_{n,b}$). Let $\mathbb{G}_n$ denote

$$\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^{N} \{f(W_i) - \mathbb{E} f(W_i)\}$$

and $\mathbb{G}_N, \mathbb{G}_n, \mathbb{G}_n$ denote the analogs for $\mathbb{G}_n$ for the other samples. Let $a \wedge b = \min\{a,b\}, a \vee b = \max(a,b)$, $a \lesssim b$ denote $a \leqslant cb$ for some $c > 0$ that does not depend on n, and $a \lesssim_P b$ to denote $a = O_P(b)$. Let $\|q\|_{F,2} := \int_{x \in \mathcal{X}} q(x)^2 dF(x)$, and $\|q\|_\infty := \sup_{x \in \mathcal{X}} |q(x)|$. Let δ_T for $\delta \in R^p, T \subset \{1,...,p\}, \delta_{T,j} := \delta_j 1_{\{j \in T\}}$. Let $\mathbf{subG}(\sigma^2)$ denote the subGaussian random variable with variance proxy σ. For a square matrix A, $\|A\| := \max \operatorname{eig}(A)$.

5.2. Tools.

Theorem 5.1 (Speed of Convergence of the Propensity Score)**.**

$$\frac{\sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)|}{\widehat{s}(x)} \lesssim_P \boldsymbol{ps}_N / s_{n,N}$$

Proof. Fix the level of probability $\delta > 0$. First, notice that $\exists N^* : \forall n \geqslant N^*$: $\mathrm{P}(\widehat{s}(x)/s_{n,N} > \underline{s}/2) \geqslant 1 - \delta$, by virtue of uniform convergence $\widehat{s}(x) \to s(x)$. On the same event, $\forall \epsilon > 0$:

$$\frac{\sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)|/s_{n,N}}{\widehat{s}(x)/s_{n,N}} \leqslant \frac{\sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)|}{\underline{s}/2} < \epsilon$$

holds. Hence, $\frac{\sup_{x \in \mathcal{X}} |\widehat{s}(x) - s(x)|}{\widehat{s}(x)} \lesssim_P \mathbf{ps}_N / s_{n,N}$. ■

Theorem 5.2 (LLN for Matrices)**.** *Let $Q_i = p_i p_i^\top \frac{\Delta_i}{s_i}$ be i.i.d symmetric non-negative kxk-matrices with $k \geqslant e^2$. Notice that $\|Q_i\| \leqslant \frac{\xi_k^2}{\underline{s} s_{n,N}}$. Then, by [10] for $\widehat{Q} = \mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{s_i}$,*

$$\mathbb{E}\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_k^2 \log N}{N s_{n,N}}}$$

As a result,

$$\|\widehat{Q}^{-1} - Q^{-1}\| \leqslant \|Q^{-1}\| \|\widehat{Q}^{-1}\| \|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_k^2 \log N}{N s_{n,N}}}$$

Let $\check{Q} := \mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{\widehat{s}_i}$ the plug-in estimator of $\widehat{Q}$.

Theorem 5.3 (Bound on the Eigenvalues of the Difference between Q and $\check{Q}$)**.**

$$
\begin{aligned}
\|\mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{\widehat{s}_i} - \mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{s_i}\| &\leqslant \sup_{x \in \mathcal{X}} \frac{|\widehat{s}(x) - s(x)|}{|\widehat{s}(x)|} \|\mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{s_i}\| \\
&\lesssim_P \frac{\boldsymbol{ps}_N}{s_{n,N}} (\|Q\| + \sqrt{\frac{\xi_k^2 \log N}{N s_{n,N}}})
\end{aligned}
$$

As a result, $\|\mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i}{\widehat{s}_i} - Q\| \lesssim_P (\frac{\boldsymbol{ps}_N}{s_{n,N}} + 1)\sqrt{\frac{\xi_k^2 \log N}{N s_{n,N}}}$

6. Proof of Theorem 3.2

Theorem 6.1 (Bound on the error term $\mathbb{E}_{n,b} p_i (\epsilon_i + r_i) \frac{\Delta_i}{s_i}$)**.**

$$
\begin{aligned}
\mathbb{E}\|\mathbb{E}_{n,b} p_i \epsilon_i \frac{\Delta_i}{s_i}\| &\lesssim_P \sqrt{\frac{k}{n}} \\
\mathbb{E}\|\mathbb{E}_{n,b} p_i r_i \frac{\Delta_i}{s_i}\| &= O(l_k c_k \sqrt{\frac{k}{n}})
\end{aligned}
$$

Proof. Step 1. Bound on $\mathbb{E}_{n,b} p_i \epsilon_i \Delta_i \frac{1}{s_i}$

$$
\begin{aligned}
\mathbb{E}\|\mathbb{E}_{n,b} p_i \epsilon_i \frac{\Delta_i}{s_i}\|^2 = \mathbb{E}\|\mathbb{E}_{n,b} p_i \epsilon_i \frac{\Delta_i}{s_i}\|^2 &= \frac{1}{N} \mathbb{E} p_i^\top p_i \epsilon_i^2 \frac{\Delta_i}{s_i^2} \\
&= O(\frac{1}{N s_{n,N}} \mathbb{E} p_i^\top p_i \frac{1}{s_i}) = O(\frac{k}{n})
\end{aligned}
$$

Step 2. Bound on $\mathbb{E}_{n,b} p_i r_i \frac{\Delta_i}{s_i}$

$$
\mathbb{E}\|\mathbb{E}_{n,b} p_i r_i \frac{\Delta_i}{s_i}\|^2 = \frac{1}{N} \mathbb{E} p_i^\top p_i r_i^2 \frac{\Delta_i}{s_i^2} = O(l_k^2 c_k^2 \frac{k}{n})
$$

∎

Theorem 6.2 (Bound on the error term $\mathbb{E}_{n,b} p_i (\epsilon_i + r_i) \Delta_i (\frac{1}{\widehat{s}_i} - \frac{1}{s_i})$)**.**

Proof. Step 1. Bound on $\|\mathbb{E}_{n,b} p_i \epsilon_i \Delta_i (\frac{1}{\widehat{s}_i} - \frac{1}{s_i})\|$

WTS:

$$
\|\mathbb{E}_{n,b} p_i \epsilon_i \Delta_i (\frac{1}{\widehat{s}_i} - \frac{1}{s_i})\|^2 \lesssim_P k/s_{n,N} (\mathbf{ps}_N / s_{n,N})^2
$$

$$
\begin{aligned}
\|\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})\|^2 &= \frac{1}{N^2}\sum_{m=1}^{k}\sum_{i,j}p_{i,k}p_{j,k}\epsilon_i\epsilon_j\Delta_i\Delta_j(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})(\frac{1}{\widehat{s}_j}-\frac{1}{s_j})\\
&\leqslant \frac{1}{N}\sum_{m=1}^{k}\sum_{i\in b}p_{i,k}^2\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})^2\epsilon_i^2\\
&= \mathbb{E}_{n,b}p_i^\top p_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})^2\epsilon_i^2\\
&\leqslant \mathbb{E}_{n,b}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2}\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\\
&\leqslant (\mathbb{E}_{n,b}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2}-\mathbb{E}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2})\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\\
&+\mathbb{E}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2}\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)},
\end{aligned}
$$

Notice that :

$$
\begin{aligned}
\mathbb{E}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2} = \mathbb{E}p_i^\top p_i\epsilon_i^2\frac{1}{s_i} &\leqslant k\|Q\|\sigma^2\frac{1}{s_{n,N}} = O(k/s_{n,N}),\\
\mathbb{E}_{n,b}p_i^\top p_i\epsilon_i^2\Delta_i\frac{1}{s_i^2}-\mathbb{E}p_i^\top p_i\epsilon_i^2\frac{1}{s_i} &\sim o_P(k/s_{n,N})\\
\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|}{\widehat{s}(x)} &\lesssim_P \mathbf{ps}_N/s_{n,N}
\end{aligned}
$$

This implies that $\|\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})\|^2 \lesssim_P k\mathbf{ps}_N^2/s_{n,N}^3$

Step 2. Bound on $\|\mathbb{E}_{n,b}p_ir_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})\|$

$$
\begin{aligned}
\|\mathbb{E}_{n,b}p_ir_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})\|^2 &= \frac{1}{N^2}\sum_{m=1}^{k}\sum_{i,j}p_{i,k}p_{j,k}r_ir_j\Delta_i\Delta_j(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})(\frac{1}{\widehat{s}_j}-\frac{1}{s_j})\\
&\leqslant \mathbb{E}_{n,b}p_i^\top p_ir_i^2\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})^2\\
&\leqslant \mathbb{E}_{n,b}p_i^\top p_ir_i^2\Delta_i\frac{1}{s_i^2}\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\\
&\leqslant \big[\mathbb{E}p_i^\top p_ir_i^2\frac{1}{s_i}\big]\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}+(\mathbb{E}_{n,b}p_i^\top p_ir_i^2\Delta_i\frac{1}{s_i^2}-\mathbb{E}p_i^\top p_ir_i^2\frac{\Delta_i}{s_i^2})\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\\
&\leqslant kl_k^2c_k^2\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\frac{1}{s_{n,N}}+(\mathbb{E}_{n,b}p_i^\top p_ir_i^2\frac{\Delta_i}{s_i^2}-\mathbb{E}p_i^\top p_ir_i^2\frac{1}{s_i})\frac{|s(x)-\widehat{s}(x)|^2}{\widehat{s}^2(x)}\\
&\lesssim_P k/s_{n,N}l_k^2c_k^2(\mathbf{ps}_N/s_{n,N})^2
\end{aligned}
$$

■

Proof of Theorem 3.1.

$$\|p^\top(\widehat{\beta}_{IPW}-\beta)\|_{F,2} \lesssim_P \|Q\|\big[\|Q^{-1}\mathbb{E}_{n,b}p_i(\epsilon_i+r_i)\frac{\Delta_i}{s_i}\| + \|Q^{-1}\mathbb{E}_{n,b}p_i(\epsilon_i+r_i)\frac{\Delta_i}{s_i}\frac{\widehat{s}_i-s_i}{\widehat{s}_i}\| \tag{6.1}$$

$$+\|Q\|^{-1}\|\check{Q}\|^{-1}\|(\check{Q}-Q)\mathbb{E}_{n,b}p_i(\epsilon_i+r_i)\frac{\Delta_i}{s_i}(\frac{-\widehat{s}_i+s_i}{\widehat{s}_i}+1)\| \tag{6.2}$$

$$\lesssim_P^i \sqrt{k/n}+\sqrt{k/s_{n,N}}(\mathbf{ps}_N/s_{n,N}) + (\sqrt{k/n}+\sqrt{k/s_{n,N}}(\mathbf{ps}_N/s_{n,N}))(\sqrt{\frac{\xi_k^2\log N}{s_{n,N}N}}) \tag{6.3}$$

Here, (i) follows from Theorems 6.1-6.2. ■

Proof of Theorem 3.2.

$$\begin{aligned}
\sqrt{n}\alpha'(\widehat{\beta}_{IPW}-\beta) &= \alpha'\check{Q}^{-1}\frac{1}{\sqrt{n}}\frac{n}{N}\sum_{i\in b}p_i(r_i+\epsilon_i)\frac{\Delta_i}{\widehat{s}_i}\\
&= \alpha' Q^{-1}\sqrt{\frac{n}{N}}\mathbb{G}_n p_i(r_i+\epsilon_i)\frac{\Delta_i}{s_i}\\
&+ \underbrace{\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})}_{\text{S1}}\\
&+ \underbrace{\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b}p_i r_i\Delta_i(\frac{1}{\widehat{s}_i}-\frac{1}{s_i})}_{\text{APX1}}\\
&+ \underbrace{\sqrt{n}\alpha'(\check{Q}^{-1}-Q^{-1})\mathbb{E}_{n,b}p_i\epsilon_i\frac{\Delta_i}{\widehat{s}_i}}_{\text{S2}}\\
&+ \underbrace{\sqrt{n}\alpha'(\check{Q}^{-1}-Q^{-1})\mathbb{E}_{n,b}p_i r_i\frac{\Delta_i}{\widehat{s}_i}}_{\text{APX2}}
\end{aligned}$$

6.0.1. *APX1.*

$$\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b}p_i r_i\frac{\Delta_i(s_i-\widehat{s}_i)}{s_i\widehat{s}_i} = O(\sup_{x\in\mathcal{X}}\frac{|s(x)-\widehat{s}(x)|}{\widehat{s}(x)}l_k c_k\sqrt{\mathbb{E}_{n,b}(\alpha' p_i)^2\frac{\Delta_i}{s_i^2}}) \leqslant \sqrt{n}\mathbf{ps}_N/s_{n,N}(l_k c_k)\frac{1}{\sqrt{s_{n,N}}}$$

6.0.2. *APX2.*

$$\begin{aligned}
\sqrt{n}\alpha'(\check{Q}^{-1}-Q^{-1})\mathbb{E}_{n,b}p_i r_i\Delta_i\frac{1}{\widehat{s}_i} &= \sqrt{n}\alpha'(\check{Q}^{-1}-Q^{-1})(\mathbb{E}_{n,b}p_i r_i\frac{\Delta_i}{s_i}+\mathbb{E}_{n,b}p_i r_i\Delta_i\frac{s_i-\widehat{s}_i}{s_i\widehat{s}_i})\\
&\leqslant \sqrt{n}\alpha'\|\check{Q}-Q\|\big[l_k c_k\sqrt{k/n}+\mathbf{ps}_N/s_{n,N}(l_k c_k)\frac{1}{\sqrt{s_{n,N}}}\big]\\
&\leqslant \sqrt{n}l_k c_k\sqrt{\frac{\xi_k^2\log N}{s_{n,N}N}}\big[\sqrt{k/n}+\mathbf{ps}_N/s_{n,N}\frac{1}{\sqrt{s_{n,N}}}\big]
\end{aligned}$$

which is true since

$$\|\check{Q}^{-1} - Q^{-1}\| = \|\check{Q}^{-1}\|\|\check{Q} - Q\|\|Q^{-1}\| \tag{6.4}$$

6.0.3. *S1.* Conditional on $(x_i)_{i=1}^N, (\Delta_i)_{i\in a}$,

$$\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i(\frac{1}{\widehat{s}_i} - \frac{1}{s_i})$$

is mean zero and has variance

$$n\text{Var}[\alpha' Q^{-1}\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i(\frac{1}{\widehat{s}_i} - \frac{1}{s_i})]|(x_i)_{i=1}^N, (\Delta_i)_{i\in a} \leqslant \|Q^{-1}\|^2\frac{n}{N}\sum_{i\in b}\alpha' p_i p_i^\top \alpha\Delta_i$$
$$(\frac{1}{\widehat{s}_i} - \frac{1}{s_i})^2\sigma^2 = O_P((\mathbf{ps}_N/s_{n,N})^2)$$

By Assumption 3.4, this is $o_P(1)$.

6.0.4. *S2.* Conditional on $(x_i)_{i=1}^N, (\Delta_i)_{i\in a}$,

$$\sqrt{n}(\check{Q}^{-1} - Q^{-1})\mathbb{E}_{n,b}p_i\epsilon_i\Delta_i\frac{1}{\widehat{s}_i}$$

has mean zero and variance bounded by $n\alpha^\top(\check{Q}^{-1} - Q^{-1})[\sum_{i\in b}p_ip_i^\top\Delta_i\frac{1}{\widehat{s}_i^2}](\check{Q}^{-1} - Q^{-1})\alpha\sigma^2 = \text{VAR}$

$$\text{VAR} = \underbrace{\frac{n}{N}\sigma^2\alpha'[(\check{Q}^{-1} - Q^{-1})]'\left[\mathbb{E}_{n,b}p_ip_i^\top\frac{\Delta_i}{s_i^2} + o_P(\mathbf{ps}_N/s_{n,N})\right][(\check{Q}^{-1} - Q^{-1})]\alpha}_{\text{VAR1}}$$
$$= O_P(\frac{\xi_k^2\log N}{s_{n,N}N})$$

∎

6.1. **Proof of Theorem 4.1.**

Proof.

$$\frac{\sqrt{n}\alpha'}{\|\alpha'\Omega^{1/2}\|}(\widehat{\beta}_{IPW} - \beta) = \frac{\sqrt{n}/\sqrt{n}\alpha'}{\|\alpha'\Omega^{1/2}\|}Q^{-1}\mathbb{G}_n[p_i(\epsilon_i + r_i)\Delta_i\frac{1}{s_i}]$$
$$+ o_P(1) = \sum_{i\in b}w_{ni}(\epsilon_i + r_i) + o_P(1),$$

where

$$w_{n,i} = \frac{\sqrt{n}\alpha' Q^{-1}}{n\|\alpha'\Omega^{1/2}\|}\frac{\Delta_i p_i}{s_i}, |w_{n,i}| \lesssim_P \frac{\sqrt{n}\xi_k}{ns_{n,N}\underline{s}} \lesssim_P \frac{\xi_k}{\sqrt{n}}, |\epsilon_i + r_i| \leqslant |\epsilon_i| + l_kc_k$$
$$n\mathbb{E}^2|w_{n,i}| \leqslant \|Q^{-2}\|\mathbb{E}(\Delta_i\alpha'\frac{p_i}{s_i})^2/(\alpha'\Omega\alpha) \lesssim_P 1$$

Now we verify Lindberg's condition for the CLT. First, by construction, we have

$$\operatorname{var}(\sum_{i\in b} w_{ni}(\epsilon_i + r_i)) = 1$$

Second, for any $\Delta > 0$,

$$\begin{aligned}
\sum_{i\in b} \mathbb{E}|w_{n,i}|^2(\epsilon_i + r_i)^2 1|w_{n,i}(\epsilon_i + r_i) > \Delta| &\leqslant^{(i)} 2n\mathbb{E}|w_{n,i}|^2\epsilon_i^2 1\{|\epsilon_i| + l_k^2 c_k^2 > \Delta/|w_{n,i}|\} \\
&+ 2n\mathbb{E}|w_{n,i}|^2 l_k^2 c_k^2 1\{|\epsilon_i| + l_k c_k > \Delta/|w_{n,i}|\} \\
&\leqslant^{(ii)} 2n\mathbb{E}|w_{n,i}|^2 \sup_{x\in\mathcal{X}} \mathbb{E}(\epsilon_i^2 1\{|\epsilon_i| + l_k c_k > c\Delta\sqrt{n}/\xi_k\}|x_i = x) \\
&+ 2n\mathbb{E}|w_{n,i}|^2 l_k^2 c_k^2 \sup_{x\in\mathcal{X}} \mathrm{P}(|\epsilon_i| + l_k c_k > c\Delta\sqrt{n}/\xi_k\}|x_i = x) \\
&\lesssim_P^{(iii)} o_P(1) + l_k^2 c_k^2 \frac{\sigma^2}{(c\Delta\sqrt{n}/\xi_k - l_k c_k)^2} \\
&\lesssim_P^{(iv)} o_P(1),
\end{aligned}$$

where (i) follows from Cauchy inequality, (iii) by Assumption 4.7. ■

7. Proof of Theorem 3.4

Proof.

$$\begin{aligned}
\sqrt{n}\alpha'(\widehat{\beta} - \beta) &= \alpha'(\mathbb{E}_{n,b} p_i p_i^\top)^{-1}\mathbb{E}_{n,b} p_i\big[(\frac{\Delta_i}{s_i}(y_i - \widehat{g}_i) + \widehat{g}_i - p_i^\top\beta)\big] \\
&= \alpha' Q^{-1}\sqrt{n}/\sqrt{n}\mathbb{G}_n p_i\big[(\frac{\Delta_i}{s_i}(y_i - g_i) + g_i - p_i^\top\beta)\big] \\
&+ \underbrace{\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b} p_i\big[(\frac{\Delta_i}{\widehat{s}_i}(y_i - \widehat{g}_i) + \widehat{g}_i - p_i^\top\beta) - (\frac{\Delta_i}{s_i}(y_i - g_i) + g_i - p_i^\top\beta)\big]}_{T_2} \\
&+ \underbrace{\sqrt{n}\alpha'\big[(\mathbb{E}_{n,b} p_i p_i^\top)^{-1} - Q^{-1}\big]\big[\mathbb{E}_{n,b} p_i\big[(\frac{\Delta_i}{s_i}(y_i - g_i) + g_i - p_i^\top\beta)\big]\big]}_{T_3} \\
&+ \underbrace{\sqrt{n}\alpha'\big[(\mathbb{E}_{n,b} p_i p_i^\top)^{-1} - Q^{-1}\big]\big[\mathbb{E}_{n,b} p_i\big[(\frac{\Delta_i}{\widehat{s}_i}(y_i - \widehat{g}_i) + \widehat{g}_i - p_i^\top\beta) - (\frac{\Delta_i}{s_i}(y_i - g_i) + g_i - p_i^\top\beta)\big]}_{T_4}
\end{aligned}$$

$$
\begin{aligned}
T_2 = &\underbrace{\alpha' Q^{-1}\mathbb{E}_{n,b} p_i (1 - \frac{\Delta_i}{s_i})(\widehat{g}_i - g_i)}_{T_{2,1}} \\
&+ \underbrace{\alpha' Q^{-1}\mathbb{E}_{n,b} p_i (\frac{\Delta_i}{s_i \widehat{s}_i})(\widehat{s}_i - s_i)(y_i - g_i)}_{T_{2,2}} \\
&+ \underbrace{\alpha' Q^{-1}\mathbb{E}_{n,b} p_i (\frac{\Delta_i}{s_i \widehat{s}_i})(\widehat{s}_i - s_i)(\widehat{g}_i - g_i)}_{T_{2,3}}
\end{aligned}
$$

Consider the term $T_{2,1}$. Conditionally on the training sample and $\{x_1, x_2, ..., x_n\}$, this term has mean zero and variance, bounded as follows:

$$
\begin{aligned}
ET_{2,1}^2 | \cup_{i\in a} z_i &\leqslant \alpha' Q^{-1}\mathbb{E}_{n,b} p_i p_i^\top (1 - \frac{\Delta_i}{s_i})^2 (\widehat{g}_i - g_i)^2 Q^{-1}\alpha \\
&\leqslant \frac{n}{N}\alpha' Q^{-1}\big[\mathbb{E} p_i p_i^\top \frac{1 - s_i}{s_i}(\widehat{g}_i - g_i)^2 + \frac{\xi_k^2 \sup_x |\widehat{g}_i - g_i| \log N}{N}\big] Q^{-1}\alpha \\
&\leqslant o_P(\frac{n}{s_{n,N} N}\xi_k^2 \mathbb{E}(\widehat{g}_i - g_i)^2) \\
&= o_P(\mathbf{g}_n^2 \xi_k^2)
\end{aligned}
$$

Assuming the $n^{-1/4}$ rate of mean square convergence of $\widehat{g}(x)$, we get $ET_{2,1} = o_P(1)$.

Consider the term $T_{2,2}$. Conditionally on $(x_i)_{i=1}^N$ and the training sample, this term has mean zero and variance, bounded as follows:

$$
\begin{aligned}
ET_{2,2}^2 | (u_i, x_i)_{i=1}^N &\leqslant \alpha' (\mathbb{E} p_i p_i^\top)^{-1} \mathbb{E}_{n,b} p_i p_i^\top \frac{\Delta_i (s_i - \widehat{s}_i)^2}{s_i^4} (\mathbb{E} p_i p_i^\top)^{-1}\alpha \\
&\leqslant \sup_{x\in\mathcal{X}} |(\widehat{s}(x) - s(x))/\widehat{s}(x)|^2 \big[o_P(\sqrt{\frac{\xi_k^2 \log N}{s_{n,N} N}}) + \alpha' (\mathbb{E} p_i p_i^\top)^{-1}\alpha\big] \\
&= o_P((\mathbf{ps}_N)^2)
\end{aligned}
$$

which is $o_P(1)$ by Assumption 3.4

Consider the term $T_{2,3}$.

$$
\begin{aligned}
\sqrt{n}\alpha' Q^{-1}\mathbb{E}_{n,b} p_i \frac{\Delta_i}{s_i \widehat{s}_i}(\widehat{s}_i - s_i)(\widehat{g}_i - g_i) &\leqslant \sqrt{n \mathbb{E}_{n,b} \alpha' Q^{-1} p_i p_i^\top Q^{-1}\alpha \frac{\Delta_i}{s_i^2 \widehat{s}_i^2}(\widehat{s}_i - s_i)^2} \sqrt{\mathbb{E}_{n,b}(\widehat{g}_i - g_i)^2} \\
&\leqslant (1 + \sqrt{\frac{\xi_k^2 \log N}{N s_{n,N}}}) \mathbf{ps}_N \mathbf{g}_n
\end{aligned}
$$

Consider the term T_3. Conditionally on $(x_i)_{i=1}^N$, this term has mean zero and variance, bounded as follows:

$$
\begin{aligned}
ET_3^2|(x_i)_{i=1}^N &\leqslant \frac{n}{N}\alpha'((\mathbb{E}_{n,b}p_ip_i^\top)^{-1} - (\mathbb{E}p_ip_i^\top))\mathbb{E}_{n,b}p_ip_i^\top(\frac{\bar{\sigma}^2}{s_i^2} + r_i^2)(\mathbb{E}_{n,b}p_ip_i^\top)^{-1} - (\mathbb{E}p_ip_i^\top))\alpha \\
&\leqslant \alpha'\|(\mathbb{E}_{n,b}p_ip_i^\top)^{-1}\|^2\|(\mathbb{E}p_ip_i^\top)^{-1}\|^2\mathbb{E}_{n,b}p_ip_i^\top - \mathbb{E}p_ip_i^\top\|^2\mathbb{E}_{n,b}\alpha'p_ip_i^\top(\frac{\bar{\sigma}^2}{s_i^2} + r_i^2)\alpha \\
&\leqslant (\frac{\xi_k^2\log N}{N})(l_k^2c_k^2 + \frac{\bar{\sigma}^2}{s_{n,N}})
\end{aligned}
$$

Consider the term T_4.

$$
T_4 = ((\mathbb{E}_{n,b}p_ip_i^\top)^{-1} - Q^{-1})QT_2 \lesssim_P o_P(\frac{\xi_k^2\log N}{N})(\sqrt{\mathbf{g}_n^2\xi_k^2 + (\mathbf{ps}_N)^2})
$$

■

Proof of Theorem 3.5.

$$
\begin{aligned}
\sqrt{n}\alpha'(\widehat{\beta}_{LR} - \beta) &= \sqrt{n}\alpha'Q^{-1}\mathbb{E}_{n,b}p_i\big[(\frac{\Delta_i}{s_i}(y_i - g_i) + g_i - p_i^\top\beta)\big] + o_P(1) \\
&= \sum_{i=1}^N w_{n,i}(\epsilon_i\frac{\Delta_i}{s_i} + r_i) + o_P(1)
\end{aligned}
$$

where

$$
w_{n,i} = \sqrt{n}/n\frac{\alpha'Q^{-1}}{\|\alpha'\Omega^{1/2}\|}p_i
$$

$$
|w_{n,i}| \lesssim_P \frac{\xi_k\sqrt{n}}{n}, |\frac{\Delta_i\epsilon_i}{s_i} + r_i| \leqslant \frac{|\epsilon_i|}{\underline{s}s_{n,N}} + l_kc_k
$$

$$
n\mathbb{E}^2|w_{n,i}| \leqslant \|Q\|^{-2}\mathbb{E}(\alpha'p_i)^2/(\alpha'\Omega\alpha) \lesssim_P 1
$$

$$
n\mathbb{E}^2|w_{n,i}\frac{\Delta_i}{s_i}| \leqslant \|Q\|^{-2}\mathbb{E}(\alpha'p_i)^2/(\alpha'\Omega\alpha) \lesssim_P 1
$$

Now we verify Lindberg's condition for the CLT. First, by construction, we have

$$
\mathrm{var}(\sum_{i\in b} w_{n,i}(\frac{\Delta_i\epsilon_i}{s_i} + r_i)) = 1
$$

Second, for any $d > 0$,

$$
\begin{aligned}
\sum_{i \in b} \mathbb{E}|w_{n,i}|^2 (\frac{\Delta_i \epsilon_i}{s_i} + r_i)^2 \mathbb{1}\{|w_{n,i}(\frac{\Delta_i \epsilon_i}{s_i} + r_i)| > d\} &\leqslant^{(i)} 2n\mathbb{E}|w_{n,i}|^2 \frac{\Delta_i \epsilon_i^2}{s_i^2} \mathbb{1}\{(\frac{\Delta_i |\epsilon_i|}{s_i} + l_k c_k) > d/|w_{n,i}|\} \\
&+ 2n\mathbb{E}|w_{n,i}|^2 l_k^2 c_k^2 \mathbb{1}\{(\frac{\Delta_i |\epsilon_i|}{s_i} + l_k c_k) > d/|w_{n,i}|\} \\
&\leqslant^{(ii)} 2n\mathbb{E}|w_{n,i}|^2 \frac{\Delta_i}{s_i^2} \sup_{x \in \mathcal{X}} \mathbb{E}\epsilon_i^2 \mathbb{1}\{(\frac{\Delta_i |\epsilon_i| s_{n,N}}{s_i} + l_k c_k s_{n,N}) > d s_{n,N} n/\sqrt{n}\xi_k\} | x_i = x \\
&+ 2n\mathbb{E} l_k^2 c_k^2 |w_{n,i}|^2 \sup_{x \in \mathcal{X}} \mathrm{P}((\frac{\Delta_i |\epsilon_i| s_{n,N}}{s_i} + l_k c_k s_{n,N}) > d s_{n,N} n/\sqrt{n}\xi_k) \\
&\lesssim_P^{(iii)} o_P(1) + l_k^2 c_k^2 \frac{\sigma^2}{(d s_{n,N} n/\sqrt{n} - l_k c_k s_{n,N})^2} \lesssim_P o_P(1)
\end{aligned}
$$

where (i) follows from Cauchy inequality, (ii) is by Assumption 4.7. ■

Proof of Lemma 4.1.

$$
\begin{aligned}
\mathbb{E}(\widehat{s}_a(x) - s(x))^2 &= \mathbb{E}(L(x^\top \widehat{\gamma}_a) - L(x^\top \gamma))^2 && (7.1) \\
&\leqslant \mathbb{E} \sup_{x \in \mathcal{X}} |L(x^\top \gamma)(1 - L(x^\top \gamma))| \| \| x^\top (\widehat{\gamma}_a - \gamma) \| && (7.2) \\
&\leqslant \mathbb{E} x' x \| \widehat{\gamma}_a - \gamma \|_1 && (7.3) \\
&\leqslant \frac{\lambda s_\gamma}{n \kappa_c^2} && (7.4)
\end{aligned}
$$

■

Proof of Lemma 4.2 . The proof of this lemma consists of three steps. First, we show that the design in 4.2 produces a matrix $\mathbb{E}_N \frac{\Delta_i}{\widehat{s}_i} x_i x_i'$ that has its sparse $s \log N$ eigenvalues bounded away from above and below, as well as bounded condition number. Second, we proof the rate of consistency of [12]. Finally, we show that the lasso prediction is consistent, which completes the proof.

Definition 7.1 (Restricted eigenvalue)**.** *Let the pseudo-norm* $\|\|_{\widehat{s},2,n}$ *be defined as :*

$$
\|\delta\|_{\widehat{s},2,n}^2 = \frac{1}{N} \sum_{i=1}^{N} (\frac{\Delta_i}{\widehat{s}_i} x_i' \delta)^2
$$

and the pseudo-norm $\|\|_{s,2,n}$ *be defined as :*

$$
\|\delta\|_{s,2,n}^2 = \frac{1}{N} \sum_{i=1}^{N} (\frac{\Delta_i}{s_i} x_i' \delta)^2
$$

For each of the norms, $\cdot \in \{\widehat{s}, s\}$*, define restricted sparse eigenvalues and condition number as follows:*

$$\kappa(m)^2_{\cdot} := \min_{\|\delta_{T^c}\| \leqslant m, \delta \neq 0} \frac{\|\delta\|^2_{\cdot,2,n}}{\|\delta\|^2} \tag{7.5}$$

$$\phi(m)_{\cdot} := \max_{\|\delta_{T^c}\| \leqslant m, \delta \neq 0} \frac{\|\delta\|^2_{\cdot,2,n}}{\|\delta\|^2} \tag{7.6}$$

$$\mu(m)_{\cdot} := \frac{\sqrt{\phi(m)}}{\kappa(m)} \tag{7.7}$$

Theorem 7.1 (Bounded Restricted Eigenvalues in $\|\|_{s,2,n}$)**.** *Let the Assumption 4.2 hold. Then, $(\frac{\Delta_i}{s_i} x_i)_{i=1}^n$ satisfies for $m = s \log n$:*

$$1/\kappa(m)_{\widehat{s}} \lesssim_P 1, \phi(m)_{\widehat{s}} \lesssim_P 1, \mu(m)_{\widehat{s}} \lesssim_P 1$$

Proof. First, let us show that

$$1/\kappa(m)_s \lesssim_P 1, \phi(m)_s \lesssim_P 1, \mu(m)_s \lesssim_P 1 \tag{7.8}$$

holds for the norm based on the true propensity score. Let us check the conditions of Proposition 2, (ii) [15] for this sequence. First $(\frac{\Delta_i}{s_i} x_i)_{i=1}^n$ is a sequence of i.i.d mean zero vectors that satisfies Riesz condition in [15]. Second, $\max_{i \leqslant n, j \leqslant p} \|\frac{\Delta_i}{s_i} x_{ij}\| \leqslant K_n/\underline{s} \leqslant \infty$. Let $c_*(m), c^*(m)$ be the minimum and maximum eigenvalues associated with $\mathbb{E}_n \frac{\Delta_i}{s_i^2} x_i x_i'$. Then, $\phi(m)_s \leqslant c^*(m+s) \lesssim_P 1, 1/\kappa(m)_s \leqslant 1/c_*(m+s) \lesssim_P 1$. Hence, 7.8 holds for the norm based on the true propensity score.

Next, notice that

$$\|\delta\|^2_{\widehat{s},2,n} = \|\delta\|^2_{s,2,n} + \frac{1}{N} \sum_{i=1}^{N} \frac{s_i^2 - \widehat{s}_i^2}{\widehat{s}_i^2} (x_i'\delta)^2 \tag{7.9}$$

$$\leqslant \|\delta\|^2_{s,2,n} + o_P(n^{-1/4}) \|\delta\|^2_{s,2,n} \tag{7.10}$$

$$\phi(m)_{\widehat{s}} \leqslant (1 + o_P(n^{-1/4}))\phi(m)_s \lesssim_P 1 \tag{7.11}$$

$$\|\delta\|^2_{\widehat{s},2,n} = \|\delta\|^2_{s,2,n} + \frac{1}{N} \sum_{i=1}^{N} \frac{s_i^2 - \widehat{s}_i^2}{\widehat{s}_i^2} (x_i'\delta)^2 \tag{7.12}$$

$$\geqslant \|\delta\|^2_{s,2,n} - o_P(n^{-1/4}) \|\delta\|^2_{s,2,n} \tag{7.13}$$

$$1/\kappa(m)^2_{\widehat{s}} \leqslant 1/\{\kappa(m)^2_s (1 - o_P(n^{-1/4})\} \lesssim_P 1 \tag{7.14}$$

$$\mu(m) \lesssim_P (1) \tag{7.15}$$

■

Let $\tilde{y}_i = \frac{\Delta_i}{\widehat{s}_i} y_i, \tilde{x}_i = \frac{\Delta_i}{\widehat{s}_i} x_i, \tilde{\epsilon}_i = \frac{\Delta_i}{\widehat{s}_i} \epsilon_i$. Let us apply Corollary 1 from [12] for $(\tilde{y}_i, \tilde{x}_i)$. Denote $\bar{\sigma}^2 = \mathbb{E}\tilde{\epsilon_i}^2 \leqslant \frac{1}{\underline{s}} \mathbb{E}\epsilon_i^2$. Then, $(\tilde{y}_i, \tilde{x}_i)_{i=1}^n$ satisfy conditions ASM and RSE ([12]) with the variance of the noise term being $\bar{\sigma} = \frac{1}{\underline{s}} \mathbb{E}\epsilon_i^2$ and the approximation error being $c_s^2 \leqslant \frac{1}{\underline{s}} c_s^2$. Corollary 1 implies:

$$\|\widehat{\gamma}-\gamma_0\|_{\widehat{s},2,n} \lesssim_P \bar{\sigma}\sqrt{\frac{s\log p}{n}}+c_s \tag{7.16}$$

and Lemma 7 [12] implies:

$$\|\widehat{\gamma}-\gamma_0\|_1 \lesssim_P \bar{\sigma}\sqrt{\frac{s^3\log p}{n}}+\sqrt{s}c_s+c_s^2\sqrt{n/logp} \tag{7.17}$$

Finally, let us show rate of consistency of lasso prediction.

$$\sqrt{\mathbb{E}(x_i'(\widehat{\gamma}-\gamma))^2} \leqslant \sup_{1\leqslant j\leqslant p}|x_{i,j}|_{1\leqslant j\leqslant p}\|\widehat{\gamma}-\gamma\|_1 \leqslant K_n(\bar{\sigma}\sqrt{\frac{s^3\log p}{n}}+\sqrt{s}c_s+c_s^2\sqrt{n/logp}) \lesssim_P n^{-1/4}$$

■

References

[1] H. Bang and J.Robins. *Doubly Robust Estimation in Missing Data and Causal Inference Models* Biometrics, 2005. `http://isites.harvard.edu/fs/docs/icb.topic156289.files/`

[2] A. Belloni, V. Chernozhukov, D. Chertverikov, and K.Kato. *Some New Asymptotics for Least Squares Series Estimators* Journal of Econometrics, 2013. `http://www.cemmap.ac.uk/publication/id/7035`

[3] K. Hirano, G.W. Imbens, and G. Ridder. *Estimation of Average Treatment Effects Using the Estimated Propensity Score* Econometrica, Vol71(4), 2003.

[4] M. Azizyan, A.Singh, L. Wasserman. *Density-Sensitive Semi-Supervised Inference* `https://projecteuclid.org/euclid.aos/1368018172`

[5] J.Hahn *On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects* Econometrica, Vol.66(2), 1998.

[6] J. Lafferty, L. Wasserman. *Statistical Analysis of Semi-Supervised Regression* `http://repository.cmu.edu/cgi/viewcontent.cgi?article=2031context=compsci`

[7] Sara A. van der Geer *High-Dimensional Generalised Linear Models and The Lasso* `http://arxiv.org/pdf/0804.0703.pdf`

[8] D.G. Horwitz and D.J. Thompson *A Generalisation of Sampling Without Replacement from Finite Universe* `http://www.stat.cmu.edu/ brian/905-2008/papers/Horvitz-Thompson-1952-jasa.pdf`

[9] V. Chernozhukov, J. Escanciano, H. Ichimura, W. Newey. *"Locally Robust Semiparametric Estimation"* `http://economics.mit.edu/files/11911`

[10] M.Rudelson *"Random Vectors in the Isotropic Position"* Journal of Functional Analysis, Vol.164, 1, 60-72, 1999.

[11] S. Firpo, C.Rothe *"Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions"* `http://www.christophrothe.net/papers/sdre_oct2015.pdf`

[12] A. Belloni, V.Chernozhukov *"High-Dimensional Sparse Econometric model: an Introduction"* `https://arxiv.org/abs/1106.5242`

[13] A. Belloni, V.Chernozhukov , Y. Wei *"Honest Confidence Regions for a Regression Parameter in Logistic Regression with a Large Number of Controls"* `http://www.cemmap.ac.uk/wps/cwp671313.pdf`

[14] A. Belloni, V.Chernozhukov *"Post-l_1-penalized estimators in high-dimensional linear regression models"* `https://arxiv.org/pdf/1001.0188.pdf`

[15] C.-H. Zhang and J. Huang *"The sparsity and bias of the LASSO selection in high-dimensional linear regression"* Ann. Statist.36

[16] B. Graham *"Inverse Probability Tilting for Moment Condition with Missing Data Models "* Review of Economic Studies (2012) 79

[17] J. Angrist, P. Hull, P.Pathak, and C.Walters *"Leveraging Lotteries for School Value-Added: Testing and Estimation"* `http://eml.berkeley.edu// crwalters/papers/VAM.pdf`

[18] P. Hull *"Estimating Hospital Quality with Quasi-Experimental Data"* `http://www.mit.edu/ hull/JMP.pdf`